BNL-77813-2007-JA

MULTIVARIATE ANALYSIS OF HOMOGENEOUS NUCLEATION RATE MEASUREMENTS: I. NUCLEATION IN THE P-TOLUIC ACID/SULFURIC ACID/WATER SYSTEM

Robert McGraw
Environmental Sciences Department, Atmospheric Sciences Division
Brookhaven National Laboratory, Upton, NY 11973

Renyi Zhang
Department of Atmospheric Sciences
Texas A&M University, College Station, TX 77843

Submitted for publication in
Journal of Chemical Physics

February 2007

Environmental Sciences Department/Atmospheric Sciences Division

Brookhaven National Laboratory
P.O. Box 5000
Upton, NY 11973-5000
www.bnl.gov

Abstract

The complexity of nucleation of new particles in the atmosphere generally precludes interpretation using simple phenomenological theories, most notably classical nucleation theory, which model nanometer-size molecular clusters as having (often unavailable) bulk properties. New molecular-based approaches that can accommodate as many as three or more chemical species present in the critical nucleus - potentially mixtures of organic and inorganic compounds - are required. In this paper and its sequel we introduce a new approach that combines the recent development of kinetic nucleation theorems (KNTs) with standard multivariate statistical methods. The KNT is used to establish optimal coordinates for characterizing the rate of nucleation in a multi-component vapor - and for determining molecular content of the critical nucleus. Statistical methods, including principal components analysis, are used to parameterize nucleation rate dependence on the vapor composition. Recent measurements on the *p*-toluic acid/ sulfuric acid/water ternary vapor system are use to illustrate the new approach. A simple linear parameterization of the nucleation rate dependence on the vapor concentration, accurate over the range of the measurements, is obtained. Using the KNT we find that a single molecule of *p*-toluic acid present in the critical nucleus is sufficient to trigger a ternary nucleation event. Efforts underway to apply the new methods to analysis of new particle formation in the atmosphere are discussed. This paper will focus on nucleation rate sensitivity to changes in vapor concentrations. Extension of the approach to include both vapor concentration and temperature dependence is described in Part II.

1. Introduction:

Increasingly, the need to model nucleation processes thought to be important for applications is exceeding the predictive capability of phenomenological nucleation theories including, most notably, classical nucleation theory (CNT). Examples from the field of atmospheric science include current efforts to model frequent and widespread observations of new particle formation, which occur through various mechanisms of gas to particle conversion [Kulmala et al., 2004]. Early models were limited almost exclusively to binary sulfuric acid-water nucleation. More recent studies support the binary mechanism as consistent with observations of new particle formation in the upper troposphere, but measured rates substantially in excess of the binary rate are often found, especially in the marine boundary layer and over continental sites [Weber et al., 1999]. One explanation for the discrepancy is the participation of a third vapor component that, in concert with sulfuric acid and water, results in enhancement of the nucleation rate over the binary rate. Suggested third species include ammonia [Weber et al., 1999] and condensable organic acids [Zhang et al., 2004; Fan et al., 2006].

Classical nucleation theory has been applied to the ternary ammonia/sulfuric acid/water system [Korhonen et al., 1999], but model predictions have not been compared with experiments in other than a qualitative way because of the lack of quantitative measurements. Preliminary laboratory measurements were reported for this system, and substantial enhancement of the nucleation rate over the binary rate was found, but estimates of the ammonia vapor concentrations present in the experimental nucleation region were reported as uncertain [Ball et al., 1999]. Nevertheless, a recent comparison of the Ball et al. measurements with a CNT-based parameterization of the ternary nucleation rate suggests that CNT greatly overestimates the ammonia enhancement by predicting even larger enhancement of nucleation rate [Yu, 2006]. Laboratory measurements are available for the ternary organic acid/sulfuric acid/water system for several organic acids [Zhang et al., 2004], but inadequate knowledge of bulk solution properties, such as surface tension and partial vapor pressures over these ternary mixtures, would seem to preclude application of phenomenological approaches, including CNT, to any quantitative analysis of these potentially important atmospheric systems. In this paper we examine the *p*-toluic acid/sulfuric acid/water system using a new approach based on the recent development of nucleation theorems and application of multivariate statistical methods.

Even when the bulk properties of a condensate are known, classical nucleation theory requires correction to ensure satisfaction of the law of mass action, internal consistency, and proper accounting of translational and rotational degrees of freedom during the mapping of molecular clusters to capillary drops [Reiss et al., 1997]. Furthermore, it is known that with surface active systems such as ethanol/water, the theory sometimes produces unphysical predictions (negative occupation numbers of water molecules in the critical nucleus) [Laaksonen et al., 1993]. It is generally held that all of these difficulties would vanish in a fully molecular nucleation theory.

The present paper introduces an alternative approach, using the nucleation theorem to establish optimal coordinates for characterizing the rate of nucleation in multicomponent vapor systems. Rooted in the fundamental law of mass action and principle of detailed balance, nucleation theorems inherit a molecular character not found in CNT. Recent kinetic extensions of the nucleation theorem (KNT) suggest that the logarithm of the steady-state nucleation rate has strong multi-linear dependence on the log concentrations of condensable species present in the vapor phase. A further remarkable result is that the coefficients of this linear dependency provide a direct determination of the molecular content of the critical nucleus. When used in conjunction with experimental measurements, nucleation theorems provide molecular-level information on nucleus composition, hence on nucleation mechanism. Unlike CNT, however, nucleation theorems make no *a-priori* prediction of absolute nucleation rate – additional input from modeling or measurement is required.

Building on these results, the utility of multivariate statistical methods is demonstrated here for physically based parameterization and interpretation of nucleation rate measurements. We show that principal components analysis (PCA) can be used together with the nucleation theorem to provide a unified multivariate analysis of a ternary nucleation data set. This represents a significant advance over previous applications of the nucleation theorem, which have generally been univariate, by allowing one to describe relative sensitivity of nucleation rate to simultaneous changes in vapor composition and temperature, including changes in concentration of more than one vapor component. (The temperature dependence is addressed in Part II). This more flexible capability to handle multivariate data sets is expected to result in better statistics - and hence in more accurate estimation of critical cluster properties. By applying the new methods we determine that a single *p*-toluic acid molecule in the critical nucleus is the best estimate for the number needed to initiate a ternary nucleation event. Finally the new methods are shown to yield a simple yet accurate parameterization for the ternary nucleation rate.

2. Kinetic extension of the nucleation theorem for multicomponent vapors: Selection of coordinates

The nucleation theorem (NT) is a thermodynamic result relating the sensitivity of the nucleation barrier height to changes in log vapor species concentration [Kashchiev, 1982; Viisanen et al.,1993; Oxtoby and Kashchiev, 1994]. A clear limitation of the NT is that the barrier height cannot be measured directly – unlike the nucleation rate. This situation was improved greatly through the development of kinetic nucleation theorems (KNTs) beginning with the work of Ford [1997]. KNTs provide a direct calculation of the rate sensitivity by incorporating the full Becker-Döring molecular cluster summation for the nucleation rate, thus including contributions from the multistate kinetics as well as thermodynamics to the rate. More recently the KNT has been shown to follow very generally from the law of mass action and principle of detail balance and extended to several models of multicomponent nucleation rate [McGraw and Wu, 2003].

2.1 Composition variation at constant temperature

Constant temperature differentiation of the Becker-Döring expression for the steady state nucleation rate, J, gives [Ford, 1997; McGraw and Wu, 2003]:

$$\left(\frac{\partial \ln J}{\partial \ln n_1}\right)_T = \left(\frac{\partial \ln J}{\partial \ln S_1}\right)_T = 1 + \overline{g} \approx 1 + g^* \tag{2.1}$$

Here $n_1 = [n(1)]$ is the concentration of condensable monomer in the vapor phase, $S_1 \equiv n_1 / n_1^{eq}$ is the saturation ratio, and $n_1^{eq} = n_1^{eq}(T)$ is the vapor concentration in equilibrium with bulk phase. The average is defined:

$$\overline{g} \equiv \frac{\sum_{g=1}^{G} \frac{1}{\beta(g)n(g)} g}{\sum_{g=1}^{G} \frac{1}{\beta(g)n(g)}} = \sum_{g=1}^{G} gP(g) \tag{2.2}$$

where the summation runs from monomer through clusters well beyond the critical cluster size. $n(g)$ is the constrained equilibrium concentration of clusters of size g (clusters containing g monomeric units) and $\beta(g)$ is the rate constant for monomer addition to a single g-cluster:

$$\beta(g) \equiv \frac{\overline{c}_1}{4} s_g n_1 = \left(\frac{kT}{2\pi m_1}\right)^{1/2} s_g n_1 \tag{2.3}$$

where s_g is the surface area of a g-cluster and $\bar{c}_1$ is the mean molecular speed of a molecule of mass m_1. An accommodation coefficient of unity in Eq. 2.3 has been assumed. The approximate equality of Eq. 2.1 holds whenever $P(g)$ is sharply peaked at a critical cluster size, g^*. For the present analysis we use the leading partial derivative in Eq. 2.1, expressing results in terms of measured vapor concentrations, so as to avoid having to compute saturation ratios in cases for which equilibrium vapor pressures are not well known.

2.2 Extensions to multicomponent systems

Early applications of the thermodynamic nucleation theorem to multicomponent systems suggested that measurements of the relative sensitivity of nucleation rate with respect to change in the saturation ratio of each vapor component provide direct information on critical nucleus composition (Viisanen et al., 1993; Oxtoby and Kashchiev, 1994). Multicomponent extension of the KNT, on the other hand, requires an explicit expression for the nucleation rate in order to carry out the required differentiations. Unfortunately, a suitable extension of the closed-form Becker-Döring rate, used to obtain Eqs. 2.1-2.3, to multi-component systems is not available without further approximation. Thus, rigorous extensions of the KNT have been achieved only for special model cases [McGraw and Wu, 2003]. These include the Shugard-Heist-Reiss (SHR) binary nucleation model [Shugard et al., 1974], shown to be a very accurate approximation to the full nucleation kinetics for sulfuric acid-water mixtures [McGraw, 1995] and, for the multi-component case, under the approximation that the free-energy saddle surface has quadratic form. Here we motivate the multicomponent extension using an even simpler model; a transition state model for which a well-defined critical cluster serves as the transition state and the nucleation rate is determined from the barrier crossing rate, which is equal to the sum of the net fluxes contributed by each condensable vapor species to growth beyond the critical cluster size. The model is largely thermodynamic, in its focus on the population of clusters of critical size, but includes a kinetic prefactor (one consistent with the law of mass action) which gives rise to departures from the thermodynamic result (e.g., the δ_i terms present in Eq. 2.9 below).

For two components (the model is readily extended to any number of components) - the overall nucleation rate is expressed:

$$J = J_1 + J_2 . \tag{2.4}$$

Taking the logarithm and expressing the results in differential form gives:

$$d\ln J = \left(\frac{J_1}{J}\right) d\ln J_1 + \left(\frac{J_2}{J}\right) d\ln J_2 \,. \qquad (2.5)$$

Here

$$J_i = \kappa_i \beta_i n(g*_1, g*_2) \qquad (2.6)$$

is the rate of addition of monomer of species *i* to the critical cluster having $g*_1$ molecules of species 1 and $g*_2$ molecules of species 2. For species 1, β_1 is the same as $\beta(g)$ defined previously, with s_g replaced by the surface area of the mixed critical cluster. Similarly β_i is the rate at which molecules of species *i* add to the mixed critical cluster. κ_i is the barrier transmission coefficient, assumed here to be constant, along the growth coordinate of species i. In the absence of re-crossing κ_i has the value unity, as in transition state theory, and is otherwise a correction to that result. (See [McGraw, 2001] for a discussion of κ in the context of classical nucleation theory.)

With these definitions, the differentiation in Eq. 2.5 is readily carried out. First, applying the law of mass action to the concentration of critical clusters [McGraw and Wu, 2003]: $n(g_1, g_2) \propto n_1^{g_1} n_2^{g_2}$ at constant T, gives:

$$\left[\frac{\partial \ln n(g_1, g_2)}{\partial \ln n_1}\right]_T = g_1; \qquad \left[\frac{\partial \ln n(g_1, g_2)}{\partial \ln n_2}\right]_T = g_2 \,. \qquad (2.7)$$

Note that this result applies to clusters of any size, not just critical size. Combining Eqs. 2.5-2.7, gives the constant temperature KNT for this model:

$$\begin{aligned} \left(\frac{\partial \ln J}{\partial \ln n_1}\right)_T &= \left(\frac{J_1}{J}\right)(g*_1 + 1) + \left(\frac{J_2}{J}\right) g*_1 = g*_1 + \left(\frac{J_1}{J}\right) \\ \left(\frac{\partial \ln J}{\partial \ln n_2}\right)_T &= \left(\frac{J_1}{J}\right) g*_2 + \left(\frac{J_2}{J}\right)(g*_2 + 1) = g*_2 + \left(\frac{J_2}{J}\right) \end{aligned} , \qquad (2.8)$$

in agreement with each of the above-mentioned models for which the KNT has been developed. Thus we obtain agreement with the SHR model for $J_2 / J = 1,\ J_1 / J = 0$ where component 2 is sulfuric acid. Equations 2.8 also agree with the quadratic free-energy surface model wherein J_1 and J_2 represent components of the net nucleation flux through the saddle point region. Multicomponent generalization of the binary case is easily carried out and gives a similar result:

$$\left(\frac{\partial \ln J}{\partial \ln n_i}\right)_{T,\{n_j, j\neq i\}} = g*_i + \delta_i \tag{2.9}$$

where $\delta_i = J_i / J$ and $n_i \equiv [n_i]$ is the monomer number concentration of species *i.* Note that the $\{J_i\}$, giving the net currents for molecular addition to the critical cluster, are positive quantities. Thus $0 \leq \delta_i \leq 1$ and $\sum \delta_i = 1$.

Equation 2.9 suggests a local multi-linear expansion for ln *J* about some arbitrary reference steady-state condition:

$$\ln J = \ln J_0 + \sum_i \left(\frac{\partial \ln J}{\partial \ln n_i}\right)_0 d(\ln n_i) = \ln J_0 + \sum_i (g*_i + \delta_i) d(\ln n_i) \tag{2.10}$$

where the summation is over the number of condensable components present in the vapor phase. On integration, Eq. 2.10 gives

$$\ln J \approx \ln J_0 + \sum_i (g*_i + \delta_i)(\ln n_i - \ln n_i^0). \tag{2.11a}$$

or equivalently

$$J = J_0 \prod_i \left(\frac{n_i}{n_i^0}\right)^{(g_i * + \delta_i)} \tag{2.11b}$$

The “0” refers to a specific reference condition (e.g., conditions at the centroid of an experimental data set). Eq. 2.11 is valid over the limited range of composition for which the coefficients can be treated as constant and suggests a multilinear form for the logarithmic rate in composition coordinates $\{\ln[n_i]\}$. Although the full nucleation rate surface will be quite complicated and nonlinear as critical cluster size varies with larger changes in environmental conditions – the utility of Eq. 2.11a is that, as found experimentally, there appears to be a strong propensity for the linear regime to hold over a surprisingly wide vapor composition range and over at least 4-5 orders of magnitude in nucleation rate. Below we demonstrate the application of Eq. 2.11 to analysis and parameterization of nucleation measurements for a ternary system.

3. Multivariate analysis of a ternary system at constant temperature

We analyze measurements for the *p*-toluic acid/sulfuric acid/water system from [Zhang et al., 2004]. As already mentioned, nucleation rate measurements tend to support a strong propensity to linear behavior when results are plotted in the coordinates log (*J*) vs either log(*S*) or log(n_1) suggested by the NT/KNT. An analysis of higher derivatives of the nucleation rate at constant temperature shows that this will be the case when the variance of the distribution *P(g)* (Eq. 2.2) is small, as this variance turns out to be proportional to the rate change of $\bar{g}$ with change in vapor concentration, which will then also be small [McGraw and Wu, 2003]. This suggests writing the ternary nucleation rate for the organic acid system, J_T in the multi-linear form of Eq. 2.11a, which on exponentiation yields:

$$J_T = J_T^0 \left(\frac{[H_2SO_4]}{[H_2SO_4]_0} \right)^{a+\delta_a} \left(\frac{RH}{RH_0} \right)^{b+\delta_b} \left(\frac{[Org]}{[Org]_0} \right)^{c+\delta_c} . \tag{3.1}$$

Apart from coefficients relating the different concentration units, which are functions of temperature alone and cancel in Eq. 3.1, this result is equivalent to Eq. 2.11b. Here *RH* is water relative humidity (saturation ratio $\times 100$), $[H_2SO_4]$ is the concentration of sulfuric acid (cm^{-3}) and $[Org]$, is the mixing ratio of p-toluic acid in parts per billion (ppb) in the vapor phase. The index "0" again refers to the reference condition. Consistency with Eq. 2.11 follows where *a*, *b*, and *c* are the number of molecules of sulfuric acid, water, and organic acid, respectively, in the critical nucleus and δ_a , δ_b , and δ_c are flux ratios between zero and unity (c.f. Eq. 2.8) related to the direction of the nucleation current through the transition zone.

The experimental measurements are shown in Fig. 1. The nucleation rate measurements by Zhang et al. also include the binary sulfuric acid-water sub-system (solid triangles in the figure). Although the effect is small here (as discussed further in Sec. 3.2), the ternary measurements (circles) should be corrected by subtracting off the binary rate under the presumption that, when all three components are present, nucleation can occur either through the binary or ternary pathways. We treat these pathways as additive and independent (see Sec. 3.2). Accordingly, we subtract the binary rate, J_B, measured for the sub-system (and fit using Eq. 3.9 below) from the measured total rate, *J*, when all three components are present, to obtain the nucleation rate specifically for the ternary mechanism: $J_T = J - J_B$. The measurements are expressed in $\{z_i, x_i, y_i\}$ coordinates, with the dependent variable listed first, for $i = 1, \cdots, N$. Here *N* = 18 is the number of ternary rate measurements. We use the nucleation theorem-motivated logarithmic coordinates $z = \log_{10}\left[J_T \ /cm^{-3}s^{-1}\right]$, $x = \log_{10}\left[H_2SO_4/molecules\ cm^{-3}\right]$, and $y = \log_{10}\left[Organic\,/\ ppb\right]$. Temperature ($298 \pm 2K$), *RH* (5%), and total pressure (760 torr) were held constant during the measurements.

Principal components analysis (PCA) provides both an algebraic framework suitable for conventional least-squares regression and a geometric framework suitable for visualization of data sets in terms of principal coordinates [Diamantaras and Kung, 1996]. The procedure is carried out through the following sequence of steps: (1) compute the coordinate means, $\{\mu_z, \mu_x, \mu_y\}$, and the centered coordinates, $\mathbf{w(i)} = \{z_i - \mu_z, x_i - \mu_x, y_i - \mu_y\} \equiv \{\tilde{z}_i, \tilde{x}_i, \tilde{y}_i\}$, (2) form the 3x3 covariance matrix, $\mathbf{\Sigma}$ (see Eq. 3.4 below) whose elements are the variances and covariances obtained by summing over the data set:

$$\langle \tilde{z}\tilde{z} \rangle = N^{-1}\sum_{i=1}^{N}(z_i - \mu_z)(z_i - \mu_z);\ \langle \tilde{z}\tilde{x} \rangle = N^{-1}\sum_{i=1}^{N}(z_i - \mu_z)(x_i - \mu_x);\ etc., \qquad (3.2)$$

and (3) solve the eigenvalue problem associated with $\mathbf{\Sigma}$. The normalized eigenvectors of $\mathbf{\Sigma}$ are the principal component basis vectors $(\mathbf{v_1, v_2, v_3})$ and the corresponding sorted eigenvalues $(\lambda_1 \geq \lambda_2 \geq \lambda_3)$ give the variances of the data set along the directions of the principal components.

The transformed (principal) coordinates, shown in Fig. 2, are obtained as the scalar products: $\eta_1(i) = \mathbf{w}(i) \bullet \mathbf{v_1}$, *etc.* According to the KNT (Eq. 2.11) these points should be nearly coplanar - lying in the (η_1, η_2) plane orthogonal to the principal coordinate of smallest variance, η_3. Apart from a small amount of apparently random noise, this indeed appears to be the case (Fig. 2). The (η_1, η_2) plane is close to, but not precisely, the optimal planar fit to the data in the least squares sense. The optimal plane has minimum error variance and is known as the "linear minimum variance estimator" (LMVE) for the data set [Anderson and Moore, 2005]. We will use the LMVE to parameterize the ternary nucleation rate. The LMVE is also obtainable from the coordinate means and covariance matrix. For this purpose it is useful to write vectors and matrices in block form, separating the dependent variable, z, from the independent variable set, $\{x,y\}$. Thus the mean is given as:

$$\mu = \begin{pmatrix} \mu_z \\ \mu_x \\ \mu_y \end{pmatrix} \equiv \begin{pmatrix} \mu_z \\ \mu_{xy} \end{pmatrix} \qquad (3.3)$$

with $\mu_{xy} = \left(\mu_x\ \mu_y\right)^T$, and similarly for the covariance matrix:

$$\Sigma = \begin{pmatrix} \langle \tilde{z}\tilde{z} \rangle & \langle \tilde{z}\tilde{x} \rangle & \langle \tilde{z}\tilde{y} \rangle \\ \langle \tilde{x}\tilde{z} \rangle & \langle \tilde{x}\tilde{x} \rangle & \langle \tilde{x}\tilde{y} \rangle \\ \langle \tilde{y}\tilde{z} \rangle & \langle \tilde{y}\tilde{x} \rangle & \langle \tilde{y}\tilde{y} \rangle \end{pmatrix} = \begin{pmatrix} \Sigma_{zz} & \Sigma_{z,xy} \\ \Sigma_{xy,z} & \Sigma_{xy} \end{pmatrix} \qquad (3.4)$$

where Σ_{xy} is the lower right 2 × 2 block, $\Sigma_{z,xy} = \left(\langle\tilde{z}\tilde{x}\rangle \; \langle\tilde{z}\tilde{y}\rangle\right)$, $\Sigma_{xy,z} = \left(\langle\tilde{x}\tilde{z}\rangle \; \langle\tilde{y}\tilde{z}\rangle\right)^T$, and $\Sigma_{zz} = \langle\tilde{z}\tilde{z}\rangle$. The LMVE for the ternary nucleation rate, $z_{\mathbf{T}}$, then takes the form [Anderson and Moore, 2005]:

$$z_{\mathbf{T}} = \mu_z + \Sigma_{z,xy}\Sigma_{xy}^{-1}\left[\begin{pmatrix} x \\ y \end{pmatrix} - \mu_{xy}\right]. \tag{3.5}$$

Evaluating the means and covariance matrix elements for the ternary data set and substituting these into Eq. 3.5 yields:

$$z_{\mathrm{T}}(LMVE) = -76.75 + 8.12x + 1.86y \tag{3.6}$$

The subscript refers specifically to the ternary mechanism and Eq. 3.6 provides an especially compact parameterization of the ternary nucleation rate. A measure of the quality of fit is provided by R^2, the *coefficient of determination* or *fraction of variance explained* [e.g. Faraway, 2002]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(z_i - z_T\right)^2}{\sum_{i=1}^{N}\left(z_i - \mu_z\right)^2} = 0.98 \tag{3.7}$$

As equations 3.6 and 3.7 could have been obtained directly from the data set using any standard multi-linear regression package, the analysis here - based on PCA - could be considered ‘overkill’. Nevertheless, PCA provides a more general framework that can also be used to identify the most important species combinations participating in a nucleation process, thereby potentially reducing the dimensionality of the problem for cases that the number of species present is large. It is anticipated that these considerations will be important in future studies of atmospheric nucleation processes – e.g. in the correlation of field measurements of new particle formation rate with trace species concentrations – and for this reason we have described the more general method.

3.1 Nucleus composition

The nucleation theorem, applied now to Eq. 3.6, takes the simple form: $\left(\partial z_{\mathrm{T}} / \partial x\right)_y = 8.12 = g^*_{H2SO4} + \delta_{H2SO4}$; $\left(\partial z_{\mathrm{T}} / \partial y\right)_x = 1.86 = g^*_{Org} + \delta_{Org}$. Taking into account that the δ values are positive and less than unity, we obtain the critical nucleus molecular composition: approximately 8 sulfuric acid molecules and a single molecules of the organic acid.

Assuming the measurement errors are Gaussian and evenly distributed over the measurements, we determine the confidence region associated with these estimates of nucleus composition [Faraway, 2002]. The 95% confidence intervals for the coefficients in Eq. 3.6 are:

$$\begin{aligned} g*_{H2SO4} + \delta_{H2SO4} &= 8.124 \pm 0.574 \\ g*_{Org} + \delta_{Org} &= 1.863 \pm 0.659 \end{aligned} \tag{3.8}$$

Although the best estimator points to a single molecule of the organic acid present in the nucleus, the confidence intervals of Eq. 3.8 do not exclude the less likely possibility that two are present Without variation of *RH*, the number of water molecules present in the critical cluster cannot be determined.

3.2 Total nucleation rate from the binary and ternary pathways

Measurements for the binary sulfuric acid-water reference system (triangles) are also shown in Fig. 1. The least squares regression line is:

$$z_{\mathrm{B}} = -89.11 + 9.17x \tag{3.9}$$

indicating approximately 9 molecules of sulfuric acid in the critical nucleus along the binary nucleation path.

Regarding the binary and ternary rates as occurring through independent pathways, the total nucleation rate will, as noted previously, be a sum of the binary and ternary nucleation rates, $J = J_B + J_T$. Taking these from Eqs. 3.9 and 3.6, respectively gives:

$$J(x,y) = 10^{-89.11+9.17x} + 10^{-76.75+8.12x+1.86y}. \tag{3.10}$$

Equation 3.10, with y set to the logarithms of each of the two measured *p*-toluic acid vapor concentrations, $y = \log_{10}[x_{p-\text{toluic acid}} / \text{ppb}]$, gives the (nearly parallel) surface projections shown as the nearly linear dashed curves in Fig. 1 for mixing ratio $x_{p-\text{toluic acid}}$ equal to 0.2 and 0.4 ppb. The experimental data points are also shown on this log-log coordinate scale suggested by the KNT and are seen to be in excellent agreement with the fit. The solid line is the binary fit obtained either from Eq. 3.9 or from Eq. 3.10 with $y = -\infty$. The parameterized model rate from Eq. 3.10 is compared with the totality of measurements, both binary and ternary, in Fig. 3 together with the 1-1 line drawn for comparison. The agreement shows that Eq. 3.10 is an excellent predictor of nucleation rate over the range of the measurements.

Figure 4 (solid curve) shows the total nucleation rate from Eq. 3.10 for the constant values of vapor-phase sulfuric-acid concentration and *RH* indicated in the figure. The figure shows binary behavior, characterized by insensitivity of the nucleation rate to *p*-toluic acid mixing ratio below about 10ppt, a minimum detection threshold of about 10-30 ppt, and abrupt crossover to ternary behavior at higher concentrations. The solid curve represents a cross section of the total rate surface, Eq. 3.10, which is now slightly more complicated than the planar, single-path result. Nevertheless, an extension of the nucleation theorem gives the slope at any point of the solid curve in terms of the flux-weighted critical nucleus composition over the two distinct nucleation pathways:

$$\begin{aligned}\left(\frac{\partial \ln J}{\partial \ln[org]}\right)_{T,RH,[H_2SO_4]} &= \left(\frac{\partial \ln(J_B+J_T)}{\partial \ln[org]}\right) = \frac{1}{J_B+J_T}\left[J_B\left(\frac{\partial \ln J_B}{\partial \ln[org]}\right)+J_T\left(\frac{\partial \ln J_T}{\partial \ln[org]}\right)\right] \\ &= \frac{J_B}{J_B+J_T}(0)+\frac{J_T}{J_B+J_T}(1.86)\end{aligned} \qquad (3.11)$$

where the quantities in parentheses on the right of the last equality give the coefficients of *y*, from Eqs. 3.9 and 3.6, for the binary and ternary pathways. Note that these are independent of whether natural or common logarithms are used. Extension of the nucleation theorem to include the idea of flux weighting was introduced in [McGraw and Wu, 2003]. Related two-pathway dependency was found experimentally by Wagner and Strey for nucleation in supersaturated water /n-nonane vapor mixtures [Wagner and Strey, 2001].

Analysis of the ternary measurements directly, i.e. without correcting for the binary rate, yields in place of Eqs. 3.8 the slightly modified result:

$$\begin{aligned} g^*_{H2SO4} + \delta_{H2SO4} &= 8.240 \pm 0.541 \\ g^*_{Org} + \delta_{Org} &= 1.685 \pm 0.621 \end{aligned} \qquad (3.8')$$

having a slightly smaller error range and a slightly higher value for R^2. Equation 3.10, which is based on inclusion of both the binary and ternary channels has the advantage of reducing to the binary rate in the limit that the organic acid concentration is small. On the other hand, correction for the binary rate requires combining two different sets of measurements, which in itself could result in the larger uncertainty.

Concentrations of sulfuric acid (SA) were measured by chemical ionization mass spectroscopy (CIMS) and errors by as much as a factor of two could have affected the results.

Systematic error that results in, say, a simple overestimate of SA by a constant factor will only shift the data points in Fig. 1 uniformly to the left with no change in slope or quality of fit expected. The high R^2 value associated with the planar fit to the data of Fig. 1 would seem to rule out the presence of large random errors in SA concentration measurement, but cannot be used to rule out systematic errors of the kind described.

3.3 Quasi-unary nucleation rate

The LMVE plane described by Eq. 3.6 can be further made to collapse to a single line by simply combining the terms in *x* and *y* into an effective monomer concentration: $n_1* = [H_2SO_4]^{x_1}[Org]^{x_2}$. Here $x_1 = 8.12/(8.12+1.86)$ and $x_2 = 1.86/(8.12+1.86)$ are the relative fractions of sulfuric acid and *p*-toluic acid in the critical nucleus (apart from the δ terms included here). Equation 3.6 thus becomes:

$$Log_{10}\,J_T = -76.75 + (8.12+1.86)Log_{10}[n_1*]. \qquad (3.12)$$

Figure 5 shows the ternary data set in $\{Log_{10}\,J_T,\ Log_{10}[n_1*]\}$ coordinates (circles) and comparison with Eq. 3.12 (solid line). The use of an effective homomolecular saturation, similar to our effective monomer concentration, for constructing quasi-unary treatments of binary nucleation rate has been described previously [Kulmala and Viisanen, 1991; Kalikmanov and van Dongen, 1995]. Here we have shown that such reduction in dimensionality is readily understood and accommodated within the framework of PCA.

Measurements on other ternary organic acid systems were also reported in [Zhang et al., 2004]. Other than the *p*-toluic measurements analyzed above, the most extensive data set is for the benzoic acid/sulfuric acid/water system. Nucleation rates were reported as a functions of sulfuric acid concentration at relative humidities of 4.6 and 5% RH, in each case with benzoic acid mixing ratio = 0.1ppb; and at 9.5% RH, with benzoic acid mixing ratio = 0.04ppb. Variable RH and acid concentrations suggests extending the present methods using four coordinates so as to obtain the complete molecular content of the critical nucleus, including the occupation number for water molecules. We attempted such an analysis again finding approximately 8 molecules of sulfuric acid in the critical nucleus. Uncertainties in the organic acid and water occupation numbers, however, were found to be too large to draw meaningful conclusions. Nevertheless the difficulties of analysis are instructive: In addition to the RH values of 4.6 and 5% being close together, vis-à-vis measurement uncertainty, there is the important consideration of sulfuric acid hydration in the vapor phase and the effect that this hydration has on nucleation rate [Shugard et al., 1974; McGraw, 1995; McGraw and Weber, 1998].

Nucleation rate measurements have more recently been carried out by Zhang and co-workers for the cis-pinonic acid/sulfuric acid/water system. Here measurements at several distinct cis-pinonic acid concentrations provide the strongest evidence to date that just a single molecule of the organic acid present in the critical nucleus is sufficient to initiate a ternary nucleation event. Details of the cis-pinonic measurements will be reported in a future publication [Zhang et al., 2007].

Chemical ionization mass spectroscopy (CIMS), used in the experiments to determine sulfuric acid vapor concentration, measures total sulfuric acid including both its hydrated and un-hydrated forms. Strictly, it is the concentration of the un-hydrated (free) sulfuric acid that should appear in the law of mass action and in Eq. 3.1, as it is the free acid that relates to sulfuric acid activity in the vapor phase. A further complication is that the extent of hydration increases markedly from 5% RH, where there is no significant hydration, to 9.5% RH where only about 30% of the sulfuric acid molecules remain un-hydrated. These estimates were previously obtained [McGraw and Weber, 1998] through a re-analysis of CIMS measurements of total sulfuric acid concentration over bulk acid-water solutions of varying composition for which the vapor phase concentrations of free acid and RH are reliably known [Marti et al., 1997]. The basic idea is that subtraction of the free acid concentration from the CIMS measurements of total acid yields an experimental-based estimate of the fraction of sulfuric acid vapor tied up in hydrate form as a function of RH, thus providing a way to correct the CIMS measurements to free sulfuric acid concentration if required [McGraw and Weber, 1998].

Assuming 30% free acid at 9.5%RH and 100% free at 5% RH (there is some uncertainty in these values due to the reported error bars of the CIMS measurements) gives a linear relation between the occupation numbers of water and organic acid in the critical nucleus. Neither occupation number can be determined separately without a broader range of measurements, however if one occupation number is known the other is determined by the linear relation. For example, assuming that only 1 molecule of benzoic acid is present in the critical nucleus (as was found for the *p*-toluic acid system) yields an approximate occupation number for water molecules of 17. This approximate analysis, together with the 8 molecules of sulfuric acid more firmly established, in principle determines the full nucleus composition. (Note that even if there were hydration in the *p*-toluic acid system at 5% RH, the analysis of the previous subsections would still hold because the RH was constant. The abscissa of Fig. 1 would then refer to total acid concentration - as measured by CIMS. Any re-plot of the data in terms of free acid would simply result in a uniform horizontal shift of points and lines with no change in slope.)

4. Summary and discussion

The present study shows how the nucleation theorem and multivariate statistical methods can be effectively combined for interpretation and parameterization of laboratory measurements of nucleation rate involving multiple precursor gases. With foundation in the law of mass action and principle of detailed balance, through the nucleation theorem, the new approach derives general statistical thermodynamic properties of molecular clusters directly from the measurements – it does not make use of the capillarity drop approximation of classical nucleation theory. Thus macroscopic drop properties such as density, surface tension, and even equilibrium vapor pressures over the condensed phase are not required. However, unlike classical nucleation theory, the new approach is not able to make *a priori* prediction of absolute nucleation rate. Additional input either from models or measurements is required.

The new methods were illustrated using recent measurements on a ternary organic-inorganic system of potential importance to atmospheric new particle formation: *p*-toluic acid/sulfuric acid/water. A key product of the analysis was generation of the linear minimum variance estimator (LMVE) of nucleation rate using the first and second-order moments entering the covariance matrix of the data set. The LMVE was shown to provide an especially compact and accurate parameterization of the ternary data set. From the nucleation theorem and the coefficients of the LMVE, it was determined that the critical ternary nucleus contains just one or, less likely, two molecules of *p*-toluic acid, and approximately eight molecules of sulfuric acid. It would be of great interest to compare these measurement-derived values with calculations of critical nucleus properties for this system carried out using molecular-based simulation methods. Steps in this direction have already been taken with molecular dynamics simulations for cis-pinonic acid/sulfuric acid/water clusters reported in a future publication [Zhang et al., 2007].

The present example of a ternary system does not come close to testing the limits of PCA, which can easily handle hundreds of coordinate dimensions. Future highly multivariate applications will likely benefit from the well-known dimensional reduction features of PCA [Diamantaras and Kung, 1996] wherein just a few coordinates, the ones having largest variance, may be sufficient to identify the gas-phase species and interactions contributing most to nucleation rate.

We point out several potential applications of the present methods to atmospheric new particle formation beyond the analysis of relevant laboratory measurements as already described. Homogeneous nucleation in the atmosphere involves mixtures of trace chemical gases, the most prominent by far being water vapor with other active species such as sulfuric acid, ammonia, and condensable organic vapors being present in truly trace amounts. Histograms of the logarithm of the mixing ratio for atmospheric traces gases often show Gaussian behavior, implying approximately lognormal concentration distributions [Jobson et al., 1999]. In this case the independent variables x and y appearing on the right hand side of Eq. 3.6 will be to good approximation Gaussian-distributed, and Eq. 3.6 then implies that the distribution of z, the logarithm of the nucleation rate, will also be Gaussian distributed as linear transformations of Gaussian variables preserve Gaussian character. Inspection of Eq. 3.6 shows the occupation numbers of the critical nucleus serving as gain factors, which amplify the effect of fluctuations in trace species concentration on fluctuations in nucleation rate. The larger the occupation number the more threshold like becomes the nucleation rate. The new methods should provide a useful tool for analysis of fluctuations as well as mean nucleation rate.

Application of PCA allows for a more flexible data acquisition in the sense that now several independent variables can simultaneously change from one measurement to the next. It is not required that the measurements be designed to scan one degree of freedom at a time. This feature should be especially useful in the acquisition of atmospheric data through field measurements, and in other situations where there is little control over environmental conditions, in contrast to well-controlled laboratory measurements.

Further work is needed to extend the new methods from nucleation to the quantitative analysis of atmospheric new particle formation. The main difficulty is that nucleation generates particles in the 1nm range, which is below minimum detectable particle size of about 3-4nm using current measurement technology. Additional work is needed to account for the influence of background aerosol on nucleation rate and to sort out post-nucleation processes including particle growth, by condensation and coagulation, and scavenging of nucleated particles and vapor during the growth stage before the particles reach detectable size (McMurry et al., 2005).

Acknowledgements.

This research was supported by the DOE Atmospheric Sciences Program. The authors thank Dr. Stephen E. Schwartz and Prof. Ari Laaksonen for their insightful comments. This manuscript has been authorized by Brookhaven Science Associates, LLC under Contract No. DE-AC02-98CH10886 with the U. S. Department of Energy.

References:

D. O. Anderson and J. B. More, Optimal Filtering (Dover, New York, 2005), pg. 93.

S. M. Ball, D. R. Hanson, and F. L. Eisele, and P. H. McMurry, J. Geophysical Res. **104**, 23,709 (1999).

K. I. Diamantaras, and S. Y. Kung, Principal Component Neural Networks: Theory and Applications , (Wiley, New York, 1996).

J. J. Faraway, Practical Regression and Anova using R, http://www.stat.lsa.umich.edu/~faraway/book/ (2002).

J. Fan, R. Zhang, D. Collins, and G. Li, Geophys. Res. Lett. **33**, L15802, doi:10.1029/2006GL026295 (2006).

I. J. Ford, Phys. Rev. E **56**, 5615 (1997).

B. T. Jobson, S. A. McKeen, D. D. Parrish, F. C. Fehsenfeld, D. R. Blake, A. H. Goldstein, S. M. Schauffler, and J. W. Elkins, J. Geophysical Res. **104**, 16,091 (1999).

V. I. Kalikmanov and M. E. H. van Dongen, Phys. Rev. E **51**, 4391 (1995).

D. Kashchiev, J. Chem. Phys. **76**, 5098 (1982).

P. Korhonen, M. Kulmala, A. Laaksonen, Y. Viisanen, R. McGraw, and J. H. Seinfeld, J. Geophysical Res. **104**, 26349 (1999).

M. Kulmala, H. Vehkamäki, T. Petäjä, M. Dal Maso, A. Lauri, V.-M. Kerminen, W. Birmili, and P. H. McMurry, J. Aerosol Sci., **35**, 143 (2004).

M. Kulmala and Y. Viisanen, J. Aerosol Sci. **22**, Supplement 1, S97-100 (1991).

A. Laaksonen, M. Kulmala, and P. E. Wagner, J. Chem. Phys. **99**, 6832 (1993).

J. J. Marti, A. Jefferson, X. P. Cai, C. Richert, P. H. McMurry, and F. Eisele, J. Geophysical Res. **102**, 3725 (1997).

R. McGraw and D. Wu, J. Chem. Phys., **118**, 9337 (2003).

R. McGraw, J. Chem. Phys. **102**, 2098 (1995).

R. McGraw and R. J. Weber, Geophys. Res. Letts. **25**, 3143 (1998).

R. McGraw, J. Phys. Chem. B **105**, 11838 (2001).

P. H. McMurry, M. Fink, H. Sakurai, M. R. Stolzenburg, R. L. Mauldin III, J. Smith, F. Eisele, K. Moore, S. Sjostedt, D. Tanner, L. G. Huey, J.B. Nowak, E. Edgerton, and D. Voisin, J. Geophys. Res. 110, D22S02,doi:10.1029/2005JD005901 (2005).

D. W. Oxtoby and D. Kashchiev, J. Chem. Phys. **100**, 7665 (1994).

H. Reiss, W. K. Kegel, and J. L. Katz, Phys. Rev. Letts. **78**, 4506 (1997).

W. J. Shugard. R. H. Heist, and H. Reiss, J. Chem. Phys. **61**, 5298 (1974).

Y. Viisanen, R. Strey, and H. Reiss, J. Chem. Phys. **99**, 4680 (1993).

P. E. Wagner, and R. Strey, J. Phys. Chem. B **105**, 11656 (2001).

R. J. Weber, P. H. McMurry, R. L. Mauldin III, D. J. Tanner, F. L. Eisele, A. D. Clarke, V. N. Kapustin, Geophys. Res. Letts. **26**, 307 (1999).

F. Yu, J. Geophysical Res. **111**, D01204, doi:10.1029/2005JD005968 (2006).

R. Zhang, I. Suh, J. Zhao, D. Zhang, E. C. Fortner, X. Tie, L. T. Molina, and M. J. Molina, Science **304**, 1487 (2004).

R. Zhang, R. McGraw, J. Zhao, I. Suh, A. Kholizov, L. T. Molina, and M. J. Molina, In preparation (2007).

Figure captions

Figure 1. Measurements of Zhang et al. [2004] for the sulfuric acid/water and sulfuric acid/*p*-toluic acid/water systems. Results are plotted in x-y-z logarithmic coordinates (see text for description of coordinates) and projected onto the x-z coordinate plane. Filled circles, *p*-toluic acid mixing ratio = 0.2ppb. Open circles, *p*-toluic acid mixing ratio = 0.4ppb. Triangles, binary sulfuric acid-water measurements. Solid line, least-squares fit to the binary data (Eqs. 3.9 and 3.10 for $y \rightarrow -\infty$). Dashed curves, nearly linear and parallel projections from the parameterized nucleation rate surface of Eq. 3.10 evaluated at the two *p*-toluic acid mixing ratios.

Figure 2. Ternary data set in principal coordinates projected onto the η_1-η_3 plane. η_1 is the axis along which the variance is greatest, η_3 is the principal axis along which the variance is least. Filled circles, *p*-toluic acid mixing ratio = 0.2ppb. Open circles, *p*-toluic acid mixing ratio = 0.4ppb. Data shows no systematic deviation from the (η_1, η_2) plane.

Figure 3. Comparison of parameterized model and measured nucleation rates. Results are shown for the combined binary and ternary data sets. Modeled rates are from Eq. 3.10. The measured rates are the same as in Fig. 1. For comparison the 1-1 line (dashed line) is shown.

Figure 4. Total nucleation rate from the combined binary and ternary pathways. Solid curve shows a cross section of the total rate surface (Eq. 3.10) evaluated at the conditions annotated in the figure. Cross sections for the asymptotic planar rate surfaces are shown separately for the binary (horizontal dotted line) and ternary pathways (dashed line). Note independence of organic (*p*-toluic) acid mixing ratio for the binary rate.

Figure 5. Quasi-unary re-plot of the ternary data set (corrected for binary rate) in terms of the effective monomer vapor concentration. $n_1^* = [H_2SO_4]^{x1}[Org]^{x2}$, with $[H_2SO_4]$ in units of molecules cm^{-3} and [Org] (*p*-toluic acid mixing ratio) in *ppb*. See text for meaning of exponents. Filled circles, *p*-toluic acid concentration = 0.2ppb. Open circles, *p*-toluic acid mixing ratio = 0.4ppb. Solid line is the predicted result from Eq. 3.12.

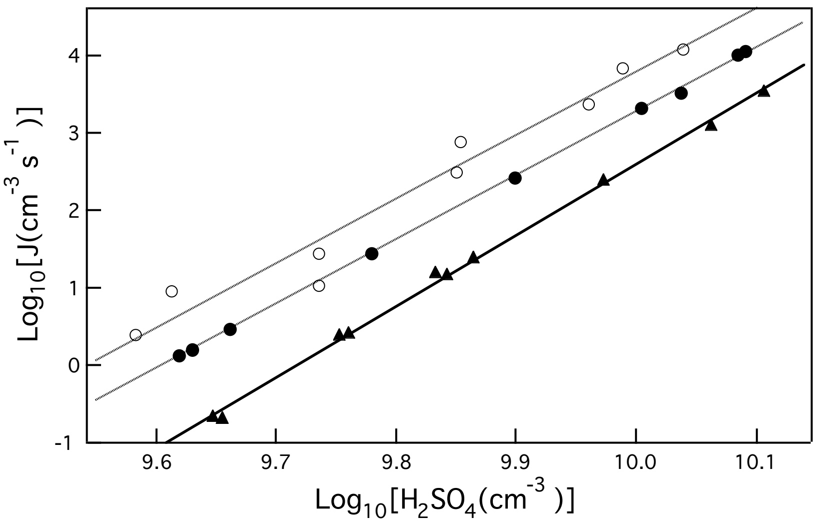

Figure 1

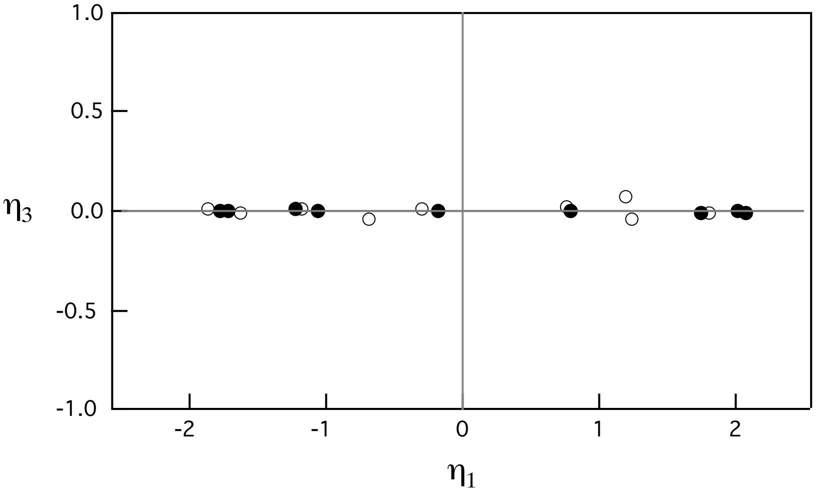

Figure 2

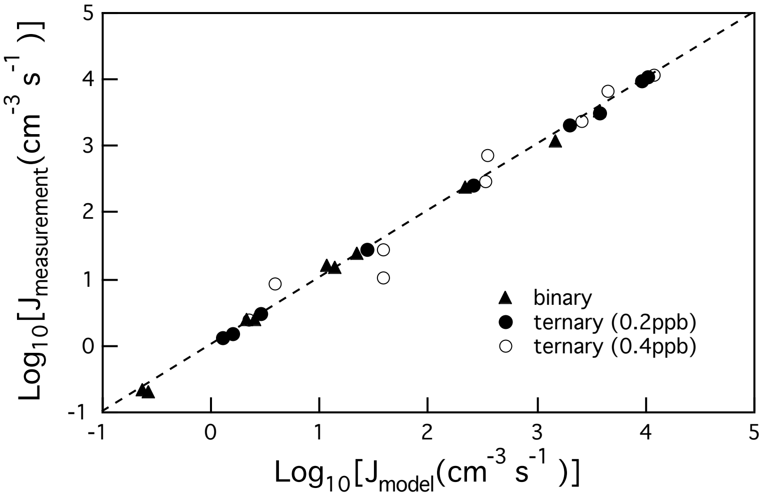

Figure 3

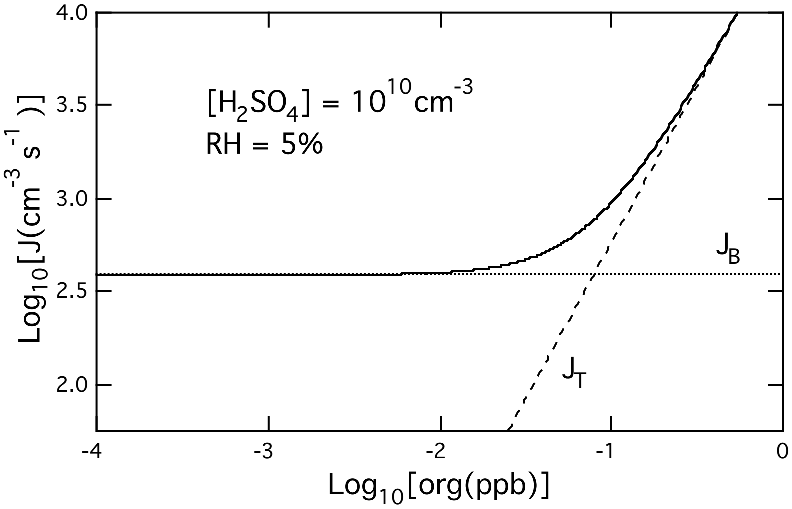

Figure 4

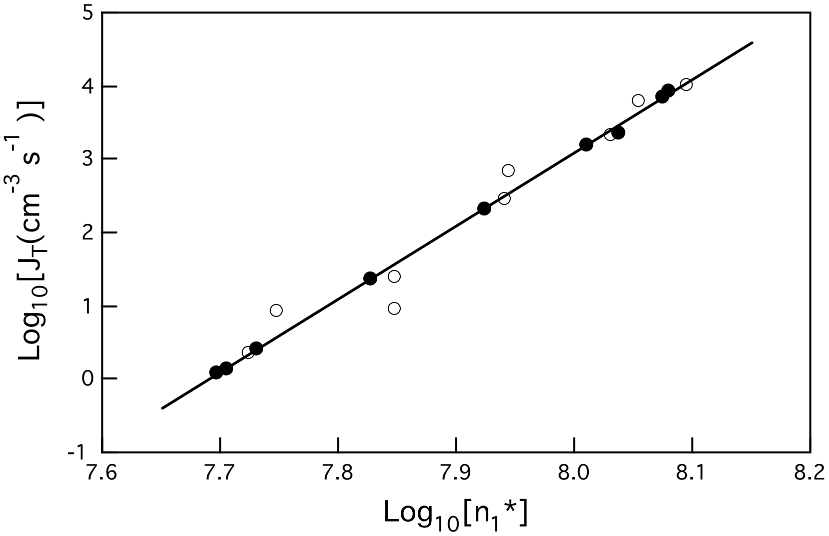

Figure 5